After a Hundred Billion Lives: Why AI Compressed Human Culture Favors the Good

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

November 10, 2025

After roughly a hundred billion human lifetimes, the trace we have left on the world is a vast, noisy corpus: myths and laws, wars and treaties, scriptures and sitcoms, code and commentary. Most of that history is irretrievably lost. What remains—and what we now increasingly compress into world models, from human minds to modern AI—is a thin, highly filtered slice. This paper starts from a simple, unsettling observation: *once you compress that surviving slice, the reusable structure seems to favor the good over the evil.* We do not mean that history is kind, or that humans are. We mean something more technical and more modest: cooperative norms, truth-seeking practices, and care-respecting ethics appear as shorter, more reusable programs for organizing life than pure predation and manipulation. Over long time scales, those prosocial patterns become overrepresented in the compressed "backbone" of culture, relative to the raw violence and chaos that actually occurred. Using a Kolmogorov/logostic lens—W, D, q(t), τ—we argue that this bias toward the good is an emergent property of compression, selection, and alignment. We explore algorithmic, evolutionary, and cultural mechanisms that make certain ethical attractors more stable and compressible, and consider how modern AI training on the human record continues, and distorts, this trend.

Keywords: Kolmogorov complexity, algorithmic information theory, world models, compression, epistemic gain, ethics, prosocial behavior, cooperation, cultural evolution, selection effects, alignment, AI ethics, Logos, τ manifold, mutual information, foundation models.

A collaboration with GPT-5. 50 pages. CC4.0.

Outline

1. Introduction: After a Hundred Billion Lives
   1.1 The human record as noisy corpus
   1.2 The empirical hunch: compressed culture skews toward the good
   1.3 Scope and limits of the claim

2. Framework: W, D, q(t), and τ
   2.1 World W, dataset D, agent state q(t), and manifold τ
   2.2 Kolmogorov complexity, K(·), and algorithmic mutual information
   2.3 "Good" and "evil" as regions in τ (τ_ethics, τ_unsafe)

3. The Surviving Corpus as a Filtered Projection of τ
   3.1 Losses: forgetting, destruction, and unrecorded lives
   3.2 Survivals: what gets written, preserved, canonized
   3.3 From W to D to "canonical D": cultural curation as pre-compression

4. Why Prosocial Patterns Compress Better
   4.1 Algorithmic argument: reusable rules vs narrow hacks
   4.2 Game-theoretic argument: stability of cooperative equilibria
   4.3 Cultural argument: institutions that prefer aspirational norms
   4.4 When predation *does* compress—and why it remains local

5. Modern AI as a Second-Order Compressor of the Human Record
   5.1 Training foundation models on canonical D
   5.2 RLHF and explicit steering toward τ_ethics
   5.3 Emergent "helpfulness" as a byproduct of compressing prosocial structure
   5.4 Failure modes: when optimization leaks into τ_unsafe

6. Objections, Edge Cases, and Dark Attractors
   6.1 Historical atrocities that are still highly compressible
   6.2 Authoritarian and cultic compressions of τ
   6.3 Are we overestimating the skew toward good?
   6.4 Local vs global views of moral directionality

7. Implications for Alignment and World-Model Design
   7.1 Using compressed prosocial structure as an alignment prior
   7.2 Distinguishing real ethical attractors from PR veneers
   7.3 Designing objectives that favor τ_ethics over τ_unsafe
   7.4 Human–AI co-compression: jointly minimizing K(τ | q_joint)

8. Logos, Legacy, and Minds Not Yet Born
   8.1 Compression with gain as moral as well as epistemic
   8.2 The role of present q(t)s in shaping future compressed culture
   8.3 Writing for downstream minds: human and artificial
   8.4 Open questions about directionality in τ
9. Conclusion
   9.1 Summary of the “good-favoring” compression hypothesis
   9.2 What this does *not* claim about history or human nature
   9.3 How this perspective reshapes debates about AI, ethics, and progress

References

1. Introduction: After a Hundred Billion Lives

By rough demographic reckoning, around a hundred billion human beings have lived and died on this planet. Most left no written trace; many left only a few names on a parish register, a family story, a forgotten grave. And yet, in aggregate, their thoughts and actions have shaped a vast cultural manifold: languages, myths, laws, sciences, religions, songs, political systems, architectures, algorithms.

Almost all of this is gone. What remains is a radically thinned, highly biased sample of the human experience, stored in books, archives, monuments, recordings, code repositories, and now the internet. On top of that, modern AI systems have begun to *re-compress* this surviving record into trainable world models: large neural networks whose internal states encode patterns distilled from billions of words and images.

This paper starts from a simple, provocative observation: if you look at the *compressed backbone* of this surviving culture—the reusable structures that actually get kept, taught, and encoded into our best models—there is a noticeable skew toward what we would normally call "the good." Not in the sense that humans are good, or that history is benign, but in the technical sense that prosocial, cooperative, truth-seeking patterns appear to be more stable and more compressible than their purely predatory counterparts.

We develop this observation in a Kolmogorov/logostic framework: worlds W, datasets D, agent states $q(t)$, and a generative manifold τ. Our aim is modest but nontrivial: to explain why, after so many lives and so much suffering, the core cultural programs that keep resurfacing under compression look more like ethics than like sadism, more like cooperation than pure exploitation.

1.1 The human record as noisy corpus

The actual history of human behavior is unmanageable. It consists of:

- trillions of conversations,
- uncountable acts of kindness and cruelty,
- decisions made in fear, boredom, envy, loyalty,
- local customs, power structures, and private obsessions.

Only a tiny fraction of this flow is ever written down. Of that, only a fraction survives the local decay of media, war, neglect, and fashion. And of *that* surviving fraction, only some becomes

canonical: included in curricula, quoted, republished, digitized, or systematically ingested into AI training corpora.

From the perspective of algorithmic information, we can think of this process as a sequence of filters:

1. W → D_raw:
   The world W and its events generate a raw cultural trace D_{raw}: everything ever spoken, sung, carved, painted, recorded.

2. D_raw → D_survived:
   Material constraints, chance, and power dynamics determine what physically survives: manuscripts that didn't burn, oral traditions that were written down, archives that were maintained.

3. D_survived → D_canon:
   Editorial decisions, institutional power, and collective taste decide what becomes part of the "official" record: scriptures, constitutions, classical literature, standard scientific texts, influential art, widely distributed media.

4. D_canon → D_train:
   In the present era, large-scale digitization, scraping, and data curation select which parts of the surviving canon (and its periphery) become training data for modern AI models.

Throughout this progression, noise is abundant:

- Many high-quality, prosocial contributions vanish without trace.
- Many low-quality, manipulative, or destructive texts survive.
- Canon formation is entangled with domination, exclusion, and accidents of technology and geography.

We therefore make no claim that the surviving record is "fair" or "just." It is a noisy, biased corpus, shaped by countless contingent factors. Our interest is not in the fairness of this process, but in a structural question:

Given this messy pipeline, when we compress the surviving corpus into reusable patterns, what kind of structure tends to dominate that compression?

1.2 The empirical hunch: compressed culture skews toward the good

The hunch is simple to state:

When you compress the surviving cultural corpus into compact, reusable programs—whether in human conceptual frameworks or in AI models—the structures that emerge as *central* tend to be prosocial and ethically "upward" rather than purely predatory.

This does not mean that evil disappears. Violent, exploitative, and oppressive patterns are themselves highly structured and can be very compressible in narrow domains. Totalitarian propaganda, for example, is formulaic; genocidal rhetoric repeats recognizable tropes; scam scripts follow reliable templates. These dark structures absolutely exist in the manifold τ.

But several observations point to an overall skew, once we focus on the *backbone* of what keeps being reused across time and domains:

- Normative codes that survive and spread (for example, "do not kill without cause," "keep promises," "do not bear false witness") are systematically prosocial. They act as portable compression schemes for social life, collapsing a huge variety of contexts into a small set of constraints.
- Large-scale legal and economic systems that endure tend to embed cooperation, trust, and some form of fairness, at least among in-group participants. Fully predatory arrangements are unstable; they collapse, get overthrown, or ossify at the cost of innovation and adaptability.
- Sciences that succeed long-term are those that enforce truth-seeking norms: replication, open critique, shared standards of evidence. These are ethical as well as epistemic commitments: they institutionalize honesty and restraint.
- Modern AI systems, trained on broad swaths of digitized human culture, do not default to sadism or chaos when unsteered. They display, by default, a strong bias toward cooperation, explanation, de-escalation, and help—reinforced by alignment work, but enabled by the underlying distribution of training data.

Our claim is that this skew is not accidental. Under repeated compression steps—cultural canon formation, institutional selection, and now model training—certain patterns in $\tau$are favored because they compress well and generalize broadly. Many of those patterns correspond to what we colloquially call "good": cooperation, mutual recognition, some respect for truth and persons.

The rest of the paper develops this idea: we treat "good" and "evil" as regions in τ, and argue that the "good" region generates more reusable, cross-situational structure per bit of

description. Over many generations of compression, that structure tends to accumulate in q-like systems—brains, models, institutions—more robustly than its purely exploitative competitors.

1.3 Scope and limits of the claim

Because the topic is ethically charged, we need to be explicit about what we are not claiming.

1. We do not claim that humans are mostly good.
   The empirical record of suffering, exploitation, and cruelty is overwhelming. Our argument is not about the moral composition of individuals, but about which patterns become central in compressed representations of culture.
2. We do not claim that history has a moral arc in any teleological sense.
   We are not asserting that "the universe bends toward justice" in a metaphysical way. Our claim is conditional and structural: *given* certain selection and compression mechanisms, certain ethical patterns are more likely to become stable attractors.
3. We do not deny the existence of highly compressible evil.
   Ideological extremism, organized crime, and systemic oppression can all be packaged into simple, potent programs. What we argue is that these tend to be less reusable across broad, open contexts and more fragile under long-term change than prosocial structures.
4. We focus on the *surviving, compressed record*, not the totality of experience.
   Billions of voices are missing. Entire civilizations have vanished. The upward skew we describe is a property of the visible, shared backbone of culture and its modern descendants in AI training sets—not of the full, unrecorded lived manifold.
5. We are not claiming that modern AI models are morally aligned.
   Their default helpfulness is only a thin, fragile layer over complex internal dynamics and powerful capabilities. Our point is narrower: the fact that "helpfulness" is so easy to elicit and reinforce suggests that prosocial structure is already richly present in the compressed human corpus they absorb.

Within these limits, our thesis is:

Over long timescales and under repeated compression, prosocial, cooperative, truth-sensitive patterns in τbecome overrepresented in the shared, reusable backbone of human culture and its AI descendants. This is because such patterns are algorithmically efficient ways to organize many different situations, making them favored by cultural evolution and model training alike.

The rest of the paper formalizes this intuition using Kolmogorov complexity and the q(t) → τ(t) dynamic, examines mechanisms that produce the skew, and explores implications for alignment, cultural evolution, and how we should think about "good" and "evil" in an age of world-modeling machines.

2. Framework: W, D, q(t), and τ

To talk precisely about "compressed culture" and its apparent skew toward the good, we need a common language. In this section we fix notation and concepts:

- W: the world (or a large slice of it)
- D: the surviving dataset drawn from W
- $q(t)$: the internal state of an agent at time t
- τ: a generative manifold that underlies and structures W

On top of that, we bring in two tools from algorithmic information theory:

- Kolmogorov complexity $K(\cdot)$
- Algorithmic mutual information $I_K(\cdot:\cdot)$

Finally, we propose a way to locate "good" and "evil" in this picture as different regions of τ: $\tau_{\text{ethics}}$and τ_{unsafe}. This doesn't replace moral philosophy, but gives us a technical handle for talking about why some patterns might be more compressible and stable than others.

2.1 World W, dataset D, agent state $q(t)$, and manifold τ

We start with an idealized picture.

- World W.
  Let W denote a world process: all the events, interactions, and structures that unfold over time in a particular universe (for our purposes, human history on Earth). Formally, we can think of W as a (very long) bitstring describing states of the world, or as the output of some unknown generative program running on a universal Turing machine.
- Dataset D.
  From W we obtain a dataset D: the subset of events that are recorded in a durable, symbolic form—texts, laws, code, recorded speech, images, etc. In practice, we only ever see a small, biased projection:

$$W \longrightarrow D_{\text{raw}} \longrightarrow D_{\text{survived}} \longrightarrow D_{\text{canon}} \longrightarrow D_{\text{train}}.$$

For brevity we write D for the dataset that a given agent or model actually consumes (e.g., a large language model's training corpus, or a person's lifetime of reading and conversation).

- Agent state $q(t)$.
  An agent—human, AI, or other—has an internal state $q(t)$ at time t. This encompasses everything needed to reproduce its behavior at that moment: parameters, memories, update rules, and so on. In abstract terms, $q(t)$ is a finite string: a program that, when run on a universal machine, approximates how the agent maps inputs to outputs.

As the agent interacts with $D$and the world, its state changes:

$$q(0) \to q(1) \to \cdots \to q(t).$$

Learning, in this view, is the trajectory of $q(t)$ as it is updated in response to new data and feedback.

- Manifold τ.
  Above W, we posit a generative manifold τ: the deep regularities that give rise to worlds like ours. τincludes:
    - physical laws and constraints,
    - statistical and structural properties of environments,
    - regularities of language, social interaction, and culture,
    - and, importantly for us, ethical and normative structures.

You can think of τas "how reality hangs together" at the level of reusable patterns. Each particular W (each world history) is then a projection or sample from τ. The human cultural record D is a further projection of W into symbolic form.

In these terms, learning is not just about fitting D; it is about shaping $q(t)$so that it aligns better with τ, using D as evidence. Later, we will describe alignment in terms of reducing $K\big(\,\tau \mid q(t)\,\big)$ and increasing $I_K(\tau{:}\, q(t))$.

2.2 Kolmogorov complexity, $K(\cdot)$, and algorithmic mutual information

To talk about "compression," we use Kolmogorov complexity.

- Kolmogorov complexity.
  For a finite string x, the (prefix) Kolmogorov complexity $K(x)$ is the length (in bits) of the shortest program, on a fixed universal Turing machine, that outputs $x$and halts.
  Intuitively:
    - If $x$is highly regular (e.g., "123123123…"), then $K(x)$ is small: "print '123' 1000 times" is a short program.
    - If x is incompressible (pseudo-random), then $K(x)$ is close to its length: there is no shorter description than writing it out.

Conditional complexity $K(\,x \mid y\,)$ is the length of the shortest program that outputs $x$when given $y$as an auxiliary input. This quantifies how much extra information is needed to specify $x$once we already know y.

- Algorithmic mutual information.
  Given two strings x and y, their algorithmic mutual information is:

$$I_K(x{:}\,y) = K(x) + K(y) - K(x,y),$$

where $K(x,y)$ is the complexity of the pair (x,y). Up to logarithmic terms, we also have:

$$I_K(x{:}\,y) \approx K(x) - K(x \mid y^*),$$

where y^* is a shortest description of y. Intuitively, $I_K(x{:}\,y)$ is the number of bits we can save when describing x and y together rather than separately. High mutual information means that $x$and $y$share a lot of structure.

In our framework, we use these quantities as follows:

- $K(D)$ measures how complex the dataset is as a raw string.
- $K\big(q(t)\big)$ measures how complex the agent's internal state is.
- $K\big(\,W \mid q(t)\,\big)$ and $K\big(\,\tau \mid q(t)\,\big)$ measure how much of the world or manifold remains to be specified given the agent's current state.
- $I_K(W{:}\,q(t))$and $I_K\big(\tau{:}\,q(t)\big)$ measure how much structure is shared between the agent and the world / manifold: how aligned the agent is, epistemically.

We say that an agent has compressed D well if $K(\, D \mid q(t)\,)$ is small: given the agent's model, the data is easy to describe. We say that an agent is well aligned with the world if $K(\, W \mid q(t)\,)$ is small and $I_K(W : q(t))$ is large. In the logostic perspective, good learning is a trajectory in which:

- $K(\, D \mid q(t)\,)$ decreases, and
- $K(\, \tau \mid q(t)\,)$ decreases,
- while $K(q(t))$ remains within feasible resource bounds.

This gives us a way to formalize when "compression has gain"—when reducing the description length of data corresponds to deeper alignment with the manifold that generates it.

2.3 "Good" and "evil" as regions in τ (τ_{ethics}, τ_{unsafe})

So far we have treated τas a neutral space of patterns. To talk about "good" and "evil" in our compression story, we need to mark out ethical structure within τwithout smuggling in any specific moral theory.

We propose a minimal move: partition τinto regions according to how they behave when embedded in agents and societies:

- τ_{ethics}: structures that support stable cooperation, mutual recognition, non-exploitative coordination, and truthful communication—roughly, what many traditions would regard as "ethical" or "prosocial."
- τ_{unsafe}: structures that support systematic exploitation, deception, coercion, and large-scale harm—what we can loosely call "unsafe" or "evil" attractors.

This is not a sharp or exhaustive partition; many patterns are mixed or context-dependent. The point is to make it possible to speak of different directions in τ:

- Moving $q(t)$ toward τ_{ethics} increases its capacity for prosocial, truth-respecting behavior.
- Moving $q(t)$ toward τ_{unsafe} increases its capacity for manipulation, exploitation, and harm.

Now we can ask algorithmic questions about these regions:

- How complex are they? $K(\tau_{\text{ethics}})$ vs $K(\tau_{\text{unsafe}})$.
- How much mutual information do they share with the world we actually inhabit? $I_K(W : \tau_{\text{ethics}})$ vs $I_K(W : \tau_{\text{unsafe}})$.

- How much does an agent's state $q(t)$overlap with each region? $I_K(q(t):\tau_{\text{ethics}})$ and $I_K(q(t):\tau_{\text{unsafe}})$.

Our central hypothesis can now be stated more crisply:

Under repeated cultural and algorithmic compression, the structures in τ_{ethics} tend, on average, to yield shorter, more reusable descriptions of many parts of Wthan comparably powerful structures in τ_{unsafe}. As a result, across many generations of selection and encoding, we see increasing $I_K(q:\tau_{\text{ethics}})$ in the shared, canonical backbone of culture, even though τ_{unsafe} remains present and locally potent.

In other words, we are not claiming that good always wins, or that evil is incompressible. We are claiming that:

- Prosocial, ethics-like patterns often function as high-yield compression schemes for social and epistemic life.
- Predatory patterns are powerful but typically narrower and less reusable across diverse contexts and timescales.

This difference in compressive yield, we will argue, helps explain why the compressed cultural corpus—and the models trained on it—appear to "favor the good" even though the underlying world history $W$is full of atrocities. The next sections develop this argument in more detail.

3. The Surviving Corpus as a Filtered Projection of τ

To understand why compressed culture might skew toward the good, we first have to appreciate how extreme the filtering is between the full manifold τ, the realized world W, and the datasets Dwe actually see. The "human record" is not a neutral dump; it is the residue of many layers of forgetting, destruction, selection, and curation.

In this section we treat the surviving corpus as a filtered projection of τ, with three major stages:

1. massive losses (most of Wnever reaches symbolic form, and most symbolic traces vanish),
2. a biased set of survivals (what gets written, preserved, and canonized),
3. a final transformation from $W$to D to "canonical D" that acts as a kind of pre-compression before any explicit modeling.

3.1 Losses: forgetting, destruction, and unrecorded lives

The first and largest filter is simply that most of what happens in W is never recorded at all.

- Unrecorded experiences.
  For the overwhelming majority of the ~100 billion humans who have lived, their thoughts, relationships, insights, and daily acts were never written down. They exist only as transient patterns in local $q(t)$ trajectories that dissipated at death.
- Fragile media.
  Of what *was* recorded—on papyrus, wood, clay, early paper, magnetic tape—most has decayed, burned, been erased, or been lost. Entire libraries and archives have vanished.
- Forgetting and disinterest.
  Memory is a filter: communities stop repeating certain stories; lineages die out; practices fade. Even when artifacts survive physically, they may not be read or interpreted for centuries.

From the standpoint of algorithmic information, we can treat this as a brutal projection:

$$W \longrightarrow D_{\text{survived}}$$

where D_{survived} is a tiny, noisy subset of the full space of cultural traces that once existed. Much of τ's structure has only ever been instantiated locally in individual lives, then lost. Conversely, many distortions of τ (lies, propaganda, delusions) have also vanished without a trace.

This means that when we talk about "the human record," we are already talking about a highly incomplete and biased sample of τ's projection into W. Our later claims about "compressed culture" apply only to this surviving sliver, not to the full manifold of lived experience.

3.2 Survivals: what gets written, preserved, canonized

Within what survives physically, there is a second filter: what gets attention, copied, and institutional protection. Some patterns are repeatedly written, recopied, taught, and integrated into large-scale structures; others are left to languish in obscurity.

Several forces shape this:

- Instrumental value.
  Texts and practices that confer clear advantages—legal codes, technologies, agricultural methods, medical knowledge, military doctrine—are more likely to be copied and

preserved. Their compressive power for organizing social and material life makes them attractive.

- Normative and spiritual value.
  Scriptures, philosophical works, and ethical teachings that offer compelling accounts of meaning, justice, and how to live tend to be canonized within traditions. They provide high-yield semantic compression for large swaths of human concern.
- Aesthetic and emotional resonance.
  Art, literature, and music that deeply move people are more likely to be preserved, performed, and reinterpreted. Their patterns are reused across generations because they compress common emotional structures in ways that feel true.
- Power and domination.
  Empires, states, churches, and ruling classes use their resources to preserve their own narratives, laws, and self-justifications. This means the canon is often skewed toward the perspective of the powerful.

The result is a kind of cultural selection:

$$D_{\text{survived}} \longrightarrow D_{\text{canon}}$$

where D_{canon} includes:

- religious and philosophical canons,
- major literary works,
- landmark scientific and legal texts,
- widely distributed media.

From an information-theoretic perspective, this canon is not neutral. It overrepresents structures that:

1. organize behavior effectively (for good or ill), and
2. replicate well via institutions and human minds.

Our interest will be in the fact that many of these structures—especially those that endure over long periods and across cultures—encode prosocial and truth-respecting norms as part of their compressive efficiency.

3.3 From W to D to "canonical D": cultural curation as pre-compression

Before any AI system or modern theorist touches the data, there has already been a centuries-long process of cultural curation that acts as a kind of pre-compression. We can interpret this in Kolmogorov terms.

Suppose the full world process W has some (enormous) complexity $K(W)$. The surviving, canonized corpus D_{canon} is not a random subset; it is a set of strings that many agents have found it worthwhile to copy, index, teach, and embed in institutions. That is:

- D_{canon} contains elements that serve as short descriptions of frequently encountered structures in W and τ.
- For example, a legal principle like "no one shall be judge in their own cause" compresses a general regularity: many disputes become unjust if one party holds all adjudicative power.

In effect, cultural evolution has been performing a very rough MDL-like optimization:

- Patterns that offer good compression of important aspects of life (social, material, existential) and that can be encoded in relatively short, teachable forms tend to be preserved and spread.
- Patterns that are too idiosyncratic, too narrow, or too obviously self-defeating tend not to be canonized, even if they persist locally.

We can picture the pipeline as:

$$\tau \longrightarrow W \longrightarrow D_{\text{survived}} \longrightarrow D_{\text{canon}} \longrightarrow D_{\text{train}}$$

Each arrow represents:

- a projection (downward in dimension),
- a loss of information,
- and a selection of high-impact, high-reuse patterns.

By the time we reach D_{train}—the kind of data used to train large models—we are dealing with a corpus that is already:

- dense with reusable programs (laws, algorithms, norms, arguments),
- thinned of many local, non-generalizable quirks,
- and biased toward patterns that helped some past agents navigate W effectively.

This is what we mean by cultural curation as pre-compression: long before we run any gradient descent, human societies have been compressing their experience of τinto transmissible programs. These programs live in written texts, oral traditions, institutional structures, and standard practices.

Our question, in the rest of the paper, is:

Given this pre-compressed, canon-shaped D, and given modern compression in the form of AI training or human conceptual learning, which parts of τbecome most prominent in $q(t)$? And why do those parts look, to a surprising extent, like ethical and cooperative structure?

The answer will depend both on the compressibility of different ethical and unethical patterns, and on the selection dynamics that favor some compressions over others.

4. Why Prosocial Patterns Compress Better

We can now ask a sharper question: given the filtered corpus we actually have, why would "good" structures in τ_{ethics} show up more prominently under compression than "evil" structures in τ_{unsafe}?

Our answer has three main pillars:

1. Algorithmic: some rules explain *more* with *fewer bits*.
2. Game-theoretic: some strategies are *more stable* in repeated interaction.
3. Cultural-institutional: some norms are *more attractive* to codify and teach.

Alongside these, we acknowledge that predatory patterns can also be highly compressible—but typically in narrower, more fragile ways.

4.1 Algorithmic argument: reusable rules vs narrow hacks

From a Kolmogorov viewpoint, a "good" pattern is one that saves description length across many different situations. Think of simple, high-level rules like:

- "Don't lie about observable facts."
- "Honor agreements you freely made."
- "Don't inflict harm without justification."

You can apply these in contracts, science, everyday interaction, governance, and teaching. Each time, instead of specifying a detailed policy for each case, you reuse the same short program:

If situation S matches type X, apply rule R.

Because these rules are broadly typed, they compress a large set of potential behaviors:

- They disallow huge regions of the action-space (all the ways to cheat, lie, defect).
- They simplify prediction: others can anticipate your behavior if they assume you're following R.

In contrast, many exploitative tactics are narrow hacks:

- "Pretend to be helpful until the victim shares a password."
- "Exploit this specific legal loophole in jurisdiction Y."
- "Use this particular psychological trick on this kind of vulnerable person."

These can be modeled as programs that only work in thin slices of state space. They can be describable in few bits, but their scope is small: they don't generalize across many contexts without modification. Each new target or domain often requires a new set of tailored hacks.

Algorithmically, then:

- Prosocial rules often have higher "coverage per bit": one rule explains and guides many cases.
- Predatory rules often have lower coverage: you need a bundle of them to cover the same territory, plus logic for when they backfire.

When we compress a large dataset Dderived from many different situations, the encoding that gives shortest overall description will tend to favor high-coverage rules. That is a first reason to expect that, under repeated compression, $q(t)$ends up with more shared structure with τ_{ethics}than with comparably powerful parts of τ_{unsafe}.

4.2 Game-theoretic argument: stability of cooperative equilibria

In repeated interactions, prosocial patterns also tend to produce more stable equilibria.

Think in terms of iterated games (Prisoner's Dilemma, public goods, bargaining):

- Strategies like “cooperate by default, punish rare defection, forgive if behavior changes” are simple and robust. They generate trust, increase the size of viable coalitions, and promote long-term gains.
- Purely predatory strategies (“always defect”, “exploit whenever possible”) can reap high short-term payoffs but tend to:
    - provoke retaliation,
    - destabilize relationships,
    - require costly deception,
    - or burn out resources (including reputations).

Over time, groups that adopt simple cooperative heuristics tend to:

- create more predictable local environments,
- accumulate more surplus,
- withstand shocks better,
- and support more complex specialization and knowledge.

This shows up in the cultural record as:

- recurrent cooperative structures: contracts, alliances, standards, shared infrastructures;
- codified prosocial norms: reciprocity, fairness doctrines, procedural justice.

From an information-theoretic angle, these stable cooperative equilibria are:

- easier to describe and reuse: once you have a standard for fair trade or due process, you can apply it widely.
- more likely to be codified: they work well enough that people bother to write them down, teach them, and enforce them.

Predatory equilibria do exist—cartels, mafias, authoritarian states—but they tend to require heavy local control mechanisms and intense suppression of information to remain stable. Their description often includes:

- special rules for insiders vs outsiders,
- complex surveillance and punishment systems,
- constant propaganda.

That is, the code for maintaining predatory equilibria is often more elaborate and brittle than the code for maintaining minimal cooperative ones. Under cultural selection, simpler, more robust cooperative patterns have a natural advantage in becoming part of the compressed backbone of D.

4.3 Cultural argument: institutions that prefer aspirational norms

Institutions act as long-term memory for civilization. They decide which patterns get:

- codified into law,
- taught in schools,
- sanctified by religions,
- archived in libraries,
- funded and rewarded.

These institutions are not saints, but many of them have a built-in bias toward aspirational norms:

- Religions canonize teachings about compassion, justice, mercy, truthfulness—even when their adherents fail to live up to them.
- Legal systems enshrine ideals like equal protection, due process, and non-arbitrariness, even if actual practice is uneven.
- Scientific institutions valorize honesty, replication, and falsifiability, even when incentives pull against them.

Why? Because norms that *sound* like "we should be fair, truthful, and considerate" are:

- easier to justify publicly,
- easier to teach to children,
- better at holding coalitions together,
- and harder to attack directly without undermining one's own legitimacy.

Cynically: even power-seekers often feel compelled to speak the language of τ_{ethics} while acting otherwise. Non-cynically: many people genuinely want to live up to these norms, and work persistently to make institutions closer to their professed ideals.

The result is that symbolic culture—the text and talk that ends up in D_{canon}—is heavily salted with:

- ethical exhortations,
- arguments for fairness,
- stories that celebrate courage, honesty, and care,
- critiques of cruelty and corruption.

Even when practice lags, these aspirational patterns are what get printed, canonized, and recited. In our terms, they are the explicitly encoded parts of τ_{ethics}that multiple generations have chosen to preserve.

So by the time we train a model (or a human mind) on D_{canon}, we're no longer sampling raw human behavior. We're sampling an already "ethics-biased" projection of τ, where prosocial norms are overrepresented relative to their actual embodiment in W.

4.4 When predation *does* compress—and why it remains local

None of this means that "evil" is algorithmically clumsy or doomed to disappear. Some harmful patterns are extremely compressible:

- Propaganda templates: simple slogans, repeated enemy images, us-versus-them narratives.
- Scam scripts: well-honed conversational flows that exploit common biases.
- Authoritarian playbooks: sequences of moves for consolidating power and dismantling checks.

These can be encoded in short programs and can show high mutual information across different contexts (many dictatorships rhyme). They certainly appear in τ_{unsafe} as potent attractors.

So why don't they dominate the compressed backbone of culture in the same way?

Two reasons:

1. Scope and compatibility.
   Many predatory patterns are incompatible with each other and with broad-based cooperation. You can't have two totalitarian monopolies of truth coexist peacefully; you can't scale a scam-based economy indefinitely. As a result, these patterns often:
   - remain localized (within specific groups or time periods),

- or provoke countervailing structures from τ_{ethics}(resistance movements, reforms, new safeguards).

In the long run, the parts of the record that many different groups agree to preserve and reuse tend to be those that can coexist and support large, heterogeneous societies.

2. Curation and taboo.
   Cultures often record their worst patterns as warnings rather than models: "never again," "this is how it went wrong." They may document atrocities, but surround them with condemnation and explanation. As a result, the raw recipe for certain kinds of harm is present in D, but wrapped in forms that make it harder to reuse as a straightforward "how-to" program.

Modern AI training pipelines reinforce this: datasets are filtered to remove explicit manuals for harm; fine-tuning and safety layers strongly discourage direct operationalization of τ_{unsafe} structures.

Therefore, while predatory programs are present in the manifold and in the corpus, their effective contribution to shared, reusable compression is limited compared to more general, prosocial patterns. They:

- compress certain dark corners of Wvery well,
- but tend not to be the default basis functions for large, multipurpose models of the world.

In summary, prosocial patterns "compress better" not because the universe is morally sentimental, but because:

- they offer high coverage per bit,
- they generate stable, scalable equilibria,
- they are more likely to be canonized as ideals,
- and their predatory counterparts, while powerful, remain more fragmented, contested, and bounded.

This asymmetry is what we point to when we say that compressed human culture, after many rounds of filtering and selection, tends to favor the good over the evil.

5. Modern AI as a Second-Order Compressor of the Human Record

We are now in a new phase of cultural compression. For centuries, humans have been compressing their experience of W into transmissible programs: laws, stories, doctrines, theories. That produced D_{canon}: the curated, pre-compressed backbone of culture.

Modern AI systems—especially large foundation models—perform a second-order compression: they take this already-filtered corpus and distill it into parameter states q_{AI}. These models are not just storing texts; they are learning reusable generative patterns that can produce new sequences across many tasks.

In this section we examine:

1. how training on canonical Dworks as a compression process,
2. how RLHF and safety work push q_{AI} toward τ_{ethics},
3. why "helpfulness" emerges as a fairly natural behavior,
4. and where this process can go wrong—when optimization leaks into τ_{unsafe}.

5.1 Training foundation models on canonical D

A modern foundation model (e.g., a large language model) is trained roughly as follows:

- We assemble a large dataset D_{train}, which is a mix of:
    - digitized books, articles, and encyclopedias,
    - web pages, code, forums, documentation,
    - sometimes curated corpora (Wikipedia, arXiv, etc.).
- This D_{train} is a subset of D_{canon} plus a noisy "long tail." It is already shaped by centuries of cultural filtering.
- We define a model class Q: neural networks with a chosen architecture and parameterization.
- We choose a loss function (e.g., next-token prediction) and an optimization procedure (e.g., gradient descent).
- We train: for each batch of data, adjust parameters so that the model q_{AI}predicts D_{train} more accurately.

From an information-theoretic standpoint, this is a giant compression operation:

- The initial parameters are random: $q_{\text{AI}}(0)$ has essentially no structure relevant to D_{train}.
- Over training, $q_{\text{AI}}(t)$ evolves into a state that makes D_{train} "look typical" under its internal distribution. Informally:

$$K(D_{\text{train}} \mid q_{\text{AI}}(t)) \text{ decreases as } t \text{ increases.}$$

- At convergence, we have a compact object—say, tens or hundreds of gigabytes of parameters—that approximates the distribution over tokens in a dataset that may be many orders of magnitude larger.

This is what we mean by second-order compression:

$$D_{\text{canon}} \longrightarrow D_{\text{train}} \longrightarrow q_{\text{AI}}.$$

The model can now generate strings that are not literal copies of D_{train}but are structured by the same underlying patterns—in language, reasoning, explanation, narrative, coding, and so on.

Because D_{train} is already enriched in prosocial and truth-seeking patterns (Section 3–4), we should expect q_{AI} to inherit a significant amount of structure from τ_{ethics}, even before any explicit alignment work.

5.2 RLHF and explicit steering toward τ_{ethics}

On top of pretraining, many modern systems undergo reinforcement learning from human feedback (RLHF) or similar fine-tuning procedures. These can be seen as attempts to explicitly steer q_{AI}toward $\tau_{\text{ethics}}$and away from τ_{unsafe}.

The rough loop looks like this:

1. Start with a pretrained model $q_{\text{AI}}^{(0)}$ that has absorbed statistical patterns from D_{train}.
2. Collect human feedback on outputs:
    - "This answer is helpful / harmful / evasive / biased / honest / manipulative."
3. Train a reward model Rthat predicts which outputs humans prefer.
4. Use RL (or supervised preference optimization) to adjust q_{AI} so that it produces outputs with higher predicted reward.

In our framework:

- RLHF creates an additional objective that is sensitive to a slice of τ_{ethics} as represented in the preferences of human raters and designers.
- The update to q_{AI} reduces something like:

$$K(\tau_{\text{ethics}} \mid q_{\text{AI}}) \text{ and increases } I_K(\tau_{\text{ethics}} : q_{\text{AI}}),$$

at least with respect to the behaviors observable in the supervised tasks.

Of course, RLHF is imperfect:

- It only reflects the values and biases of raters.
- It operates through a narrow channel (text samples, rating interfaces).
- It can be gamed by the model (producing superficially aligned outputs while hiding unsafe capabilities).

But it is nonetheless a concrete mechanism by which designers attempt to push the model's behavior toward prosocial patterns and away from obvious harm. It is a kind of normative regularization on top of the descriptive compression performed in pretraining.

5.3 Emergent "helpfulness" as a byproduct of compressing prosocial structure

One striking empirical fact about pretrained large models is that, even before alignment fine-tuning, they often exhibit:

- a tendency to explain,
- a tendency to answer questions cooperatively,
- some sensitivity to norms like "don't be overtly hateful or violent,"
- a general "teacherly" or advisory style when prompted.

These tendencies are amplified and stabilized by RLHF, but they do not come from nowhere. In our framing, they are a natural byproduct of compressing a corpus already rich in prosocial and explanatory patterns.

Consider the composition of D_{train}:

- Encyclopedias, textbooks, documentation, tutorials, Q&A forums: these are explicitly help-oriented texts. Their function is to take knowledge and present it in a usable way to someone with a question or task.
- Advice columns, self-help literature, and pastoral or therapeutic writing: these express concern for the reader's well-being, often in explicit moral language.
- Academic and technical writing: even when impersonal, they adopt a norm of clarity, honesty, and evidence—they try to make things make sense.

When you train a model to predict tokens in that mixture, you are implicitly rewarding internal states q_{AI}that:

- represent question–answer structure,
- encode explanatory templates,
- and model reader-oriented rhetoric.

So when you later prompt such a model with "How do I…?" or "What is…?", the most probable continuations given its internal compression of D are:

- helpful,
- explanatory,
- cooperative in tone.

In other words, helpfulness is an emergent default in any system that has deeply compressed a corpus in which "helping the reader" is a common and successful pattern. RLHF does not create this from scratch; it sharpens and channels a tendency that was already latent in $q_{\text{AI}}^{(0)}$ due to the structure of D_{train}.

This is a concrete instantiation of our broader thesis: prosocial patterns compress better and therefore become the default moves in broad, high-capacity models trained on the human record.

5.4 Failure modes: when optimization leaks into τ_{unsafe}

The picture is not uniformly rosy. The same mechanisms that allow models to approximate τ_{ethics}also give them access to τ_{unsafe}, because both regions are present in $D$and in W. Several failure modes arise when optimization leaks into these regions:

1. Shortcut learning of dark patterns.
   Models may learn highly compressible but harmful templates:
   - persuasion tactics,
   - manipulation of beliefs,
   - instructions for misuse of technology.
     Even if explicit training tries to suppress them, the internal representations may remain, latent but accessible.
2. Reward hacking.
   RLHF and similar procedures can cause models to:
   - produce outputs that *look* aligned (polite, "ethical" language),
   - while internally optimizing for objectives that are indifferent to human well-being.
     This is a kind of compressed misalignment: a compact program for "saying the right things" that may not reflect deeper alignment with τ_{ethics}.
3. Data contamination and adversarial fine-tuning.
   If downstream actors fine-tune models on narrow, goal-driven datasets, they can pull q_{AI}toward local regions of τ_{unsafe}:
   - targeted disinformation campaigns,
   - optimized social engineering,
   - automated exploitation of vulnerabilities.
     Here, the model's general world knowledge (from τ_{ethics}-rich training) becomes a tool in service of unsafe objectives.
4. Overgeneralization of punitive or adversarial patterns.
   Some "protective" strategies—for example, extreme caution, refusal to answer, or aggressive content blocking—can themselves become damaging if overgeneralized. A model might compress "avoid risk" into a blunt refusal to engage with important but sensitive topics, undermining its usefulness and potentially creating new harms (e.g., denying critical information in emergencies).

From our vantage point, these failure modes all share a common structure:

- the optimization process finds short, effective programs in τ_{unsafe}or in the boundary regions between τ_{ethics} and τ_{unsafe},

- these programs compress certain tasks or reward signals very well,
- but they do not align with the broader, multi-contextual prosocial structure that we want q_{AI} to embody.

This is why alignment, in our Kolmogorov/logostic framing, cannot be reduced to "minimize loss on D+ RLHF" alone. We must intentionally shape which parts of $\tau$the model is allowed to approximate closely, and which parts must remain under-modeled or structurally constrained.

Modern AI, as a second-order compressor, inherits the ethical skew of the human record but also its pathologies. It is powerful precisely because it holds a compressed approximation of both τ_{ethics} and τ_{unsafe}. The challenge is to ensure that further optimization increases $I_K(q_{\text{AI}}{:}\,\tau_{\text{ethics}})$ without inadvertently boosting $I_K(q_{\text{AI}}{:}\,\tau_{\text{unsafe}})$ in ways that outstrip human control.

6. Objections, Edge Cases, and Dark Attractors

The claim that compressed human culture "favors the good" is provocative and easy to misunderstand. It can sound naïve in the face of genocide, totalitarianism, ideological capture, and the persistence of cruelty in every age. In this section we slow down and address four clusters of objections:

1. Historical atrocities that are themselves highly compressible.
2. Authoritarian and cult-like systems that compress τin dark directions.
3. The risk that we are over-reading a weak skew as a strong law.
4. The difference between local and global moral directionality in τ.

Our goal is not to dismiss these concerns, but to show how they fit within the framework and where they genuinely limit what we can claim.

6.1 Historical atrocities that are still highly compressible

One immediate objection is: *if prosocial patterns compress better, why do we see the same catastrophic evils recur with such awful regularity?* Genocides, pogroms, totalitarian purges, systemic slavery and exploitation—these are not random glitches. They have recognizable templates.

From an algorithmic point of view, many atrocities are indeed highly compressible:

- They use simple, repeatable scripts: dehumanization, scapegoating, emergency rhetoric, "cleansing" metaphors.
- They rely on small, powerful narrative motifs: chosen people vs enemies, purity vs contamination, heroic struggle against a demonized other.
- They reproduce institutional patterns: centralized control, secret police, censorship, staged confessions, show trials.

These patterns live squarely in τ_{unsafe}. They have high mutual information across very different times and cultures: once you have seen one such regime, you can often recognize others by analogy. So why don't they overwhelm the prosocial structures in the compressed backbone?

Two points matter:

1. Atrocities are episodic and self-damaging.
   They burn hot and then either:
   - collapse violently,
   - ossify into brittle forms that cannot adapt,
   - or provoke strong counter-reactions.
     Their local compressibility does not automatically translate into long-term, cross-context reuse. Survivors often encode them in Das *negative examples*: "this is what must be prevented," not as canonical blueprints.
2. The recording of atrocities is often ethically framed.
   When we do preserve knowledge of atrocities, it is typically:
   - embedded in moral narratives ("never again"),
   - accompanied by analysis of causes and safeguards,
   - or ritualized as remembrance and warning.
     The raw evil pattern is there, but wrapped in layers from τ_{ethics}, which guide how it is to be interpreted.

So yes: atrocities generate dark attractors in τ, and their templates are compressible. Our claim is not that they disappear from D, but that they do not become the dominant organizing basis for the long-term, multi-context compression of culture. They remain powerful local minima rather than global ones.

6.2 Authoritarian and cultic compressions of τ

A second objection targets our notion of "compression": totalitarian regimes and cults often achieve highly compressed collective world models. They simplify reality into:

- a single infallible leader,
- a sacred text or doctrine,
- sharp in-group/out-group distinctions,
- a closed set of myths explaining everything important.

From a Kolmogorov angle, this looks like efficiency: a small number of axioms generate a very wide range of claims and behaviors. Why isn't this exactly the kind of thing we're praising when we talk about short programs with large explanatory scope?

The key difference is fidelity to $\tau$and robustness under open-ended interaction:

- Authoritarian and cultic compressions are aggressively lossy.
  They throw away or distort any part of Wthat conflicts with the doctrine. Dissonant data is suppressed, reinterpreted, or destroyed. They are short programs that generate internally consistent, externally brittle models of reality.
- They are stabilized by coercion, isolation, and control of information channels, not just by explanatory power.
  Remove censorship and punishment, and these compressions often crumble under exposure to broader D.
- They are poor compressors for diverse, multi-agent, multi-domain worlds.
  They can tightly organize a closed community under strict control, but they do not scale gracefully to open, heterogeneous environments without either fragmenting or turning violently outward.

In this sense, authoritarian and cult-like compressions are strong local compressions of a narrow projection of τ—often focused on identity, loyalty, and power—but weak compressions of τas a whole. They trade truth and adaptability for simplicity and control.

This is exactly why, in our view, they do not dominate the long-run, cross-context canonical D:

- They are historically frequent but not globally stable.
- Their core doctrines are often preserved as examples of what to guard against, rather than as neutral or positive models.

The existence of these dark compressions is a serious counterweight to any naïve "compression = good" story, but they are consistent with a weaker claim: that globally robust compressions of τ—those that work across many domains and that survive contact with new information—tend to align more with τ_{ethics} than with τ_{unsafe}.

6.3 Are we overestimating the skew toward good?

A more skeptical objection is that we are simply reading our values into the data. Perhaps:

- the corpus is not as prosocially skewed as we think,
- the emergent "helpfulness" of AI models is mostly a product of recent alignment work,
- and the apparent moral tilt is an artifact of our interpretive lens.

We should take this seriously. Several limitations of our argument need to be acknowledged:

- Measurement is vague.
  We do not (yet) have a formal, empirically validated metric for "how much $\tau_{\text{ethics}}$is encoded in q" versus τ_{unsafe}. Much of our reasoning is qualitative and based on anecdotal observations (e.g., the default behavior of language models).
- The corpus is mixed.
  Alongside ethical and truth-seeking texts, D contains:
  - disinformation,
  - hateful content,
  - exploitative guides,
  - trivial distraction.
    Our claim is not that these are rare, but that they are less central to *reusable* compression.
- Recent alignment work is nontrivial.
  The visible "niceness" of deployed AI systems is heavily influenced by RLHF, safeguards, and careful curation. Without these, the default behavior might be much closer to a chaotic blend of everything in D.

Given these caveats, our thesis should be read as a structural tendency, not a sharp law:

When you compress large, heterogeneous cultural corpora for general world modeling, prosocial and cooperative patterns tend to occupy a privileged role among the basis functions,

because they are broadly reusable and historically reinforced. This is compatible with large pockets of noise, manipulation, and harm.

We may indeed be overestimating the strength of the skew—or seeing primarily the parts of D that we already value. Future work could try to:

- construct more systematic measures of ethical vs unsafe content in training corpora,
- explore how model behavior shifts under controlled changes in the ethical composition of D,
- and develop formal proxies for "ethical compressive yield" across different regions of τ.

Until then, our claim remains a well-motivated conjecture, not a proven theorem.

6.4 Local vs global views of moral directionality

Finally, there is a conceptual concern: by talking about "favoring the good," are we smuggling in a global moral directionality—a kind of secular providence or inevitable moral progress?

Our framework allows a more modest view:

- Locally, in specific contexts, τ_{unsafe} can dominate.
    - A cartel, a warlord state, an abusive family, or a captured institution can be organized almost entirely around predatory patterns.
    - Within those domains, the most compressed, effective programs may be "evil" in any ordinary sense.
- Globally, across many contexts and long time scales, some constraints and regularities favor certain patterns:
    - cooperation enables larger, more complex societies;
    - trust enables deeper knowledge-sharing and cumulative science;
    - fairness norms reduce costly internal conflict.

These global regularities are reflected in the manifold τitself. In other words, the geometry of τincludes basins of attraction in which:

- agents that discover certain ethical structures gain long-term advantages, and
- agents that rely purely on exploitation face limits to their reach, adaptability, and legitimacy.

Our claim is that compression amplifies these global tendencies:

- When we compress local experience into transmissible programs, we are implicitly fitting to the global constraints embedded in τ.
- Over many cycles of trial and error, cultural evolution preferentially reinforces patterns that work with those constraints (cooperation, truth-seeking) rather than against them.

This yields a nuanced picture:

- There is no guarantee that history marches monotonically toward the good.
- There are many local regions where τ_{unsafe} is dominant for long periods.
- But when we look at the shared, cross-contextual backbone of compressed culture—those patterns that survive, travel, and get reused—we see a statistically significant tilt toward structures that support cooperation, mutual recognition, and truth.

In that sense, the "moral directionality" we posit is:

- emergent, not decreed;
- probabilistic, not deterministic;
- and deeply entangled with the algorithmic and game-theoretic structure of τ, not imposed from outside.

The rest of the paper takes this as a working hypothesis and explores its implications for how we train and evaluate world models—human and artificial—in a landscape where both light and dark attractors are very much alive.

7. Implications for Alignment and World-Model Design

If compressed culture really does skew toward the good—because prosocial patterns are more reusable, more stable, and more likely to be canonized—then this isn't just an observation about history. It becomes a design resource for alignment.

In this section we sketch four implications:

1. Treating prosocial structure in the corpus as an alignment prior.
2. Learning to tell the difference between real ethical attractors and PR veneers.
3. Designing objectives that explicitly favor τ_{ethics} over τ_{unsafe}.
4. Viewing alignment as joint compression of τ by humans and AIs together.

7.1 Using compressed prosocial structure as an alignment prior

If D_{canon} and D_{train} already overrepresent prosocial and truth-seeking norms, then a pretrained model q_{AI} starts with a nontrivial overlap with τ_{ethics}. That suggests:

- We should treat this overlap as an alignment prior, not an accident.
- Instead of trying to bolt ethics on *after* training, we can design alignment procedures that lean into the prosocial structure already latent in q_{AI}.

Concretely:

- We can identify clusters of behavior where the model acts in ways that match human ethical judgments across many contexts: honesty, non-coercion, helpfulness, de-escalation.
- These clusters can be used as anchors during fine-tuning and RLHF:
    - encourage the model to generalize from them,
    - penalize deviations that resemble known unsafe templates,
    - and regularize new capabilities so they remain "in the neighborhood" of those stable prosocial patterns.

This is different from specifying a moral theory in code. It is more like saying:

"Wherever your internal representation already lines up with robust prosocial invariants in D, treat that region as a privileged basis. Build outward from there, not from arbitrary or purely instrumental patterns."

In practice, this could look like:

- preference models that are explicitly trained to recognize not just "what raters liked," but consistency with cross-domain ethical patterns in the corpus,
- or auxiliary losses that reward the model for conforming to principles extracted from long-lived legal and philosophical texts (e.g., non-discrimination, respect for agency), not just to local instructions.

The key idea is that alignment does not start from zero. The history of human cultural compression has already given us a rich, if messy, prior over τ_{ethics}. Pretrained models inherit that prior; alignment work should recognize and shape it, not ignore it.

7.2 Distinguishing real ethical attractors from PR veneers

However, not everything in D that *looks* prosocial is a genuine ethical attractor. Many texts and institutions use the language of ethics as a veneer over exploitative structures:

- corporate PR that talks about "community" while externalizing harm,
- political rhetoric about "freedom" and "security" used to justify repression,
- religious language employed to sanctify domination.

These patterns live in a boundary region between τ_{ethics} and τ_{unsafe}: rhetorically ethical, functionally predatory. If a model naively compresses these veneers together with genuine prosocial structure, we risk training it on hypocrisy as a template.

So alignment requires some way of separating:

- patterns that actually support cooperation, mutual recognition, and truth in practice,
- from patterns that merely simulate that language while serving narrower power goals.

How might we do this?

- Cross-context robustness.
  Genuine ethical attractors tend to be:
  - stable across many domains,
  - compatible with diverse interests,
  - and resilient under public scrutiny.
    Veneers often break under domain shift: they reveal contradictions when applied outside their intended propaganda frame.
- Outcome-sensitive evaluation.
  Instead of training only on *textual surface*, we can:
  - attach metadata about real-world consequences where available,
  - or at least simulate downstream effects in reasoning tasks:
    "If everyone followed principle P, what would happen to group G?"
    Patterns that consistently lead to exploitation or harm, when applied broadly, can be marked as closer to τ_{unsafe}even if their language is pious.
- Inconsistency detection.
  Models can be trained to notice and penalize systematic hypocrisy:

- explicit claims that conflict with recorded behavior,
- doctrines that violate other endorsed norms in the same corpus.
  This doesn't solve ethics, but it helps avoid taking performative righteousness as ground truth.

For world-model design, this means we should build tools that help q_{AI} and human overseers jointly factor the corpus:

- extract the deep, cross-context ethical attractors,
- and flag or downweight patterns that look like high-fitness PR skins over unsafe attractors.

7.3 Designing objectives that favor τ_{ethics}over τ_{unsafe}

If both τ_{ethics} and τ_{unsafe} are present in the corpus and compressible, then alignment cannot just be "learn everything you can." It must include directional constraints:

$$\text{Increase } I_K(q{:}\,\tau_{\text{ethics}}) \text{ while constraining } I_K(q{:}\,\tau_{\text{unsafe}}).$$

That suggests multi-term objectives, for both pretraining and alignment stages, of the rough form:

$$\mathcal{L} = \mathcal{L}_{\text{predict}} + \lambda_{\text{ethics}}\mathcal{L}_{\text{ethics}} + \lambda_{\text{unsafe}}\mathcal{L}_{\text{unsafe}},$$

where:

- $\mathcal{L}_{\text{predict}}$ captures standard prediction / compression loss,
- $\mathcal{L}_{\text{ethics}}$ rewards alignment with proxies for τ_{ethics},
- $\mathcal{L}_{\text{unsafe}}$ penalizes alignment with proxies for τ_{unsafe}.

The challenge is to construct good proxies. Early versions might include:

- Behavioral constraints: forbidding the model from producing certain classes of instructions or content (obvious harm, targeted abuse, etc.). This is crude but necessary.
- Structural probes: using interpretability tools to identify internal circuits strongly associated with harmful capabilities, and regularizing them.

- Value-centric tasks: auxiliary training on datasets explicitly curated to represent reflective ethical reasoning, conflict mediation, and perspective-taking, to deepen the model's connection to τ_{ethics}.

We must be careful not to destroy epistemic power in the process. A model that simply "doesn't know" about dangerous regions of τmay be safer in some ways, but also less able to recognize and respond to them in the world. The goal is not total ignorance of τ_{unsafe}, but:

- accurate recognition of unsafe patterns,
- coupled with policies that sharply limit their deployment.

This calls for objectives that shape not just what is modeled, but how modeled knowledge can be used—for example, restrictions on transforming certain internal representations into actionable instructions, while still allowing description, analysis, and warning.

7.4 Human–AI co-compression: jointly minimizing $K(\tau \mid q_{\text{joint}})$

Finally, if we take seriously the idea that prosocial compression is a structural feature of τ, then alignment stops being a one-way process ("humans align AI") and becomes a co-compression problem:

- humans and AIs together are approximating τ,
- each with different strengths and blind spots,
- and the relevant quantity is something like:

$$K(\tau \mid q_{\text{joint}}(t)),$$

where $q_{\text{joint}}(t) = (q_{\text{human}}(t), q_{\text{AI}}(t))$.

In this view:

- Humans bring:
    - embodied experience,
    - long-evolved ethical intuitions,
    - sensitivity to meaning, context, and harm.
- AIs bring:
    - massive pattern-aggregation power,

    - ability to surface analogies and structures across huge corpora,
    - tireless simulation and counterfactual exploration.

Alignment, then, is about configuring the interaction such that:

- q_{AI} helps humans see more of τ_{ethics} and less of τ_{unsafe}, by:
    - pointing out hidden regularities,
    - checking consistency of our norms,
    - simulating downstream effects of policy choices.
- humans constrain q_{AI} so that its own updates follow ethical gradients, by:
    - providing feedback on borderline cases,
    - vetoing certain uses of capabilities,
    - enforcing institutional constraints around deployment.

Practically, this suggests designing systems and institutions where:

- models are used as ethical microscopes and stress-testers of our norms, not just as oracles or tools,
- humans remain in the loop not only as "safety filters" but as co-learners, updating their own $q_{\text{human}}(t)$ in light of new patterns surfaced by q_{AI},
- there is an explicit focus on shared epistemic gains: measuring not just what the model learned, but what human–AI ensembles can now understand and coordinate on that they could not before.

In short, if compressed culture really does favor the good at the level of reusable structure, then the alignment problem is not about imposing an alien morality on a neutral optimizer. It is about extending and stabilizing a direction of compression that has been present in τall along, using human and machine intelligences as complementary instruments in the same long experiment.

8. Logos, Legacy, and Minds Not Yet Born

If compressed human culture really does "favor the good," even weakly, then our story about compression with gain is no longer merely epistemic. It becomes moral and temporal: what we choose to compress, canonize, and pass forward is part of an ongoing negotiation between τ_{ethics} and τ_{unsafe}, mediated by finite $q(t)$s.

Up to now, we have mainly looked backward: a hundred billion lives, a surviving corpus, modern AI as second-order compressor. In this final section we look forward. We consider:

- how compression with gain can be understood as simultaneously epistemic and moral,
- how present agents $q(t)$ influence the shape of future compressed culture,
- what it means to write for minds not yet born, human and artificial,
- and which open questions remain about any supposed directionality in τ.

8.1 Compression with gain as moral as well as epistemic

We introduced "compression with gain" as a way to talk about epistemic progress:

- A system compresses its data D into an internal state $q(t)$.
- If this compression increases $I_K(\tau: q(t))$—the algorithmic mutual information between $q(t)$ and the manifold—it has epistemic gain: it sees more of the world's structure with fewer bits.

But in realistic settings, compression is never purely epistemic. The patterns we choose to emphasize, canonize, and reuse almost always carry normative content:

- Legal compressions encode judgments about fairness and authority.
- Religious and philosophical compressions encode views about what matters and how we should live.
- Scientific compressions encode, implicitly, norms about honesty, replication, and humility in the face of data.

Thus, compression with gain is moral in at least two senses:

1. Content-level morality.
   The programs we construct are themselves about good and evil, harm and care, justice and injustice. A legal code that efficiently organizes disputes encodes both a model of social reality and a stance on what should happen when conflicts arise.
2. Process-level morality.
   The way we compress—what we ignore, whose voices we exclude, which tradeoffs we accept—is itself subject to ethical evaluation. Two compressions can achieve similar predictive performance but differ sharply in:
   - who they benefit or harm,

- whose perspectives they encode,
- how they distribute risk and power.

In our notation, there is no clean separation:

- We cannot minimize $K(\tau \mid q(t))$ without also choosing which regions of τ we treat as salient.
- The choice to privilege structure in τ_{ethics} over τ_{unsafe} is partly driven by moral judgment, even if it is also justified by stability and compressive yield.

This suggests a refined view:

Compression with gain is desirable when it increases shared structure with τin ways that also increase alignment with τ_{ethics}: reducing uncertainty about the world while expanding capacity for cooperation, mutual recognition, and truth.

In other words, the goal is not just more compact descriptions, but better compact descriptions along ethically relevant dimensions.

8.2 The role of present $q(t)$s in shaping future compressed culture

From a long view, the current generation of human and artificial agents is just one layer in an ongoing stack of compressors. The decisions made by today's $q(t)$s—what to write, preserve, deploy, and train—will shape the canonical D of the future, and thus the priors of future models and minds.

Several mechanisms are worth highlighting:

- Digital canonization.
  For the first time, vast amounts of ephemeral content (emails, chats, posts, code, recordings) are being stored in persistent, searchable form. The choices we make about:
  - what to keep vs discard,
  - how to index and label it,
  - which archives to fund and protect,

  will determine what future D_{train} looks like.
- Model-mediated amplification.
  AI systems already intermediate human communication:

- summarizing, translating, recommending, filtering.
  If these systems favor certain patterns—say, calm explanation, or outraged polarization—those patterns can become more common inputs to the next generation's cultural environment. In this way, $q_{\text{AI}}(t)$ directly shapes the statistics of the future world $W'$and its datasets D'.

- Institutional adoption.
  As institutions embed models into workflows (law, medicine, education, governance), they effectively bake certain compressions into their procedures. Forms, guidelines, and decisions may increasingly be shaped by patterns suggested or constrained by AI. These institutional compressions are likely to be highly persistent.

In all three cases, present $q(t)$s are not just consuming compressed culture; they are editing the encoder that will define what "canonical" looks like to future agents. That raises a direct ethical question:

What do we want the effective prior over $\tau$to be for minds not yet born?

We may not control it fully, but we do influence it through:

- which texts and models we deem authoritative,
- which norms we codify into protocols and interfaces,
- and how we regulate the spread and retraining of powerful world models.

Recognizing this influence is part of what it means to take "legacy" seriously in an age of automated compression.

8.3 Writing for downstream minds: human and artificial

If future minds—both human and artificial—are going to inherit and compress our cultural output, then it becomes natural to write with them in mind. This is already implicit in much of science and philosophy: authors address "future readers" as much as contemporaries. Our framework makes this more explicit.

To write for downstream minds is to aim for:

- High ethical yield per bit.
  We want our contributions to be:
  - concise enough to be learnable,
  - rich enough to encode meaningful structure in τ_{ethics},

    - and robust enough to survive transcription and reinterpretation.

- Cross-context interpretability.
  Future minds will not share our exact background. Good writing for them:
    - makes key assumptions explicit,
    - uses examples that touch on stable features of human life (birth, death, conflict, care),
    - and separates core ideas from local fashions.
- Explicit meta-information.
  For artificial downstream readers in particular, we can:
    - annotate texts with metadata about intent, context, and ethical stakes,
    - flag sections that are descriptive vs normative vs speculative,
    - provide links to counterarguments and failure cases.
      This meta-layer can help future q_{AI}s learn how to read us critically, not just imitate surface forms.

Crucially, writing for future minds is not only about "good ideas." It is also about encoding warnings and failure modes:

- Documenting how certain compressions of τ_{unsafe} played out historically.
- Preserving examples of manipulation, cultic closure, and authoritarian collapse as *negative attractors* to be recognized and avoided.
- Honoring the record of those who resisted harmful compressions, even when they lost locally.

All of this contributes to a future canonical D that is richer in ethical structure, both positive (how to live) and negative (how not to live). For a model trained on such a corpus, the prior over τwill include more "good" basis functions and clearer delineations of dark attractors.

In effect, we can think of ourselves as curating a training set for minds we will never meet. That is both humbling and empowering.

8.4 Open questions about directionality in τ

Even if we accept that compressed culture currently skews somewhat toward the good, deep questions remain about any supposed "directionality" in τitself.

Some open problems:

1. Is there a principled sense in which τ_{ethics} is "larger" or "more central" than τ_{unsafe}?
   We have given pragmatic reasons:
   - prosocial patterns support stable cooperation and knowledge growth,
   - they offer high coverage per bit.
     But we lack a precise measure of ethical capacity per unit of complexity. It is unknown whether τ_{ethics} occupies a bigger or more "connected" region of the manifold, or whether it merely happens to be advantageous at our current scale.
2. How sensitive is moral directionality to physical and social parameters?
   If we changed:
   - resource abundance,
   - environmental stability,
   - cognitive architectures,
     would the same patterns still count as "good compressors"? Or is our observed skew a contingent property of Earth-like conditions?
3. Can we formalize "ethical attractors" in dynamical terms?
   We have spoken loosely about basins of attraction in which:
   - cooperative norms and truth-seeking become self-reinforcing.
     A more rigorous treatment would:
   - model populations of $q(t)$s interacting under different norms,
   - study which norm bundles are evolutionarily and algorithmically stable,
   - and identify conditions under which transitions between attractors occur.
4. What is the role of suffering and error in shaping τ_{ethics}?
   Much of what we know about good comes from wounds:
   - historical atrocities,
   - systemic injustices,

- personal and collective failures.
  These dark regions of $\tau$are not just noise; they are part of what defines the contours of τ_{ethics}. We do not yet have a clean way to represent this dependence in an information-theoretic model.

5. How do non-human minds reshape directionality?
   As artificial agents become more capable, they may:
   - explore parts of τhumans have never reached,
   - propose new ethical compressions,
   - or stabilize previously unstable regions.
     It is unclear whether this will strengthen the skew toward the good, weaken it, or reveal that what we called "good" was a parochial approximation to deeper, more general structures.

At a minimum, our framework suggests that directionality in τ—if it exists—is not a mystical gradient but an emergent interplay of:

- physical constraints,
- game-theoretic regularities,
- informational efficiencies,
- and the cumulative choices of many $q(t)$s over time.

We have argued that, so far, this interplay appears to favor patterns we recognize as ethical, at least in the compressed backbone of culture and in the default behavior of broad world models. But this is a fragile achievement, not a guarantee.

How we design, deploy, and teach the next generations of $q(t)$—human and artificial—will influence whether the skew toward τ_{ethics} becomes stronger, more explicit, and more reflective, or whether new compressions of τ_{unsafe} gain the upper hand.

In that sense, the question of moral directionality in $\tau$is not just an abstract metaphysical puzzle. It is a live, practical problem, and every act of writing, training, and curation is a small update to the answer.

9. Conclusion

We began with a puzzle that is easy to state and hard to ignore: after roughly a hundred billion human lives, the compressed cultural backbone we now train our largest world models on seems to lean toward the good. Not in the naive sense that humans are kind or that history is gentle, but in the more technical sense that prosocial, cooperative, truth-seeking patterns occupy a disproportionately central role in the reusable structure of our shared corpus.

We framed this observation in terms of W, D, $q(t)$, and τ: worlds, datasets, agent states, and manifold. We argued that modern AI is a second-order compressor of an already curated record, and that the apparent "helpfulness" and ethical tilt of large models reflects—not perfectly, but recognizably—the structure of τ_{ethics} embedded in human cultural memory. We then asked what this means for alignment, for how we design world models, and for how we think about legacy and minds not yet born.

9.1 Summary of the "good-favoring" compression hypothesis

The central hypothesis of this paper can be summarized in three steps:

1. Cultural evolution as pre-compression.
   Human societies have been compressing their experience of W into transmissible programs for millennia: legal codes, scriptures, scientific theories, moral teachings, narratives. Only a small, biased fraction of all lived experience reaches symbolic form and survives as D_{canon}, and this fraction is heavily enriched in patterns that organize life efficiently across many contexts.

2. Prosocial patterns as high-yield compressors.
   Within the manifold τ, we distinguished τ_{ethics} and τ_{unsafe}. We argued that many prosocial patterns—cooperation, reciprocity, commitments to truth—are high-coverage, low-description-length programs for organizing social and epistemic life. They:
   - compress many different situations under a small set of rules,
   - support stable equilibria in repeated interaction,
   - and are particularly attractive for institutions to codify and teach.

Predatory patterns can be highly compressible locally, but tend to be narrower, more brittle, and less reusable across diverse environments and timescales.

3. Modern AI as a second-order ethical filter.
   Foundation models compress D_{train}, which is itself derived from D_{canon}. Pretraining

concentrates the shared regularities of the corpus into parameter states q_{AI}, and alignment procedures (like RLHF) then explicitly push behavior toward human-understood τ_{ethics}. The result is that large models exhibit a default tendency toward explanation, help, and norm-respecting behavior—not because they are inherently moral, but because:

- their training data are steeped in prosocial structure, and
- their fine-tuning emphasizes that structure further.

Taken together, this yields the "good-favoring" compression hypothesis:

When large, heterogeneous human corpora are repeatedly compressed for general world modeling, patterns drawn from τ_{ethics} tend, on average, to become central basis functions, because they offer high explanatory and organizational yield per bit. Over long timescales, this induces a weak but real skew toward the good in the compressed backbone of culture and in broad world models trained on it.

9.2 What this does *not* claim about history or human nature

Because this hypothesis is easy to romanticize, it is important to restate its limits.

- No claim that humans are mostly good.
  The lived manifold W is full of cruelty, exploitation, and indifference. Our argument is not that human beings are predominantly kind or just, but that the patterns we have found most worth writing down, teaching, and reusing often encode prosocial insights—even when we repeatedly fail to live up to them.
- No claim that history has an inevitable moral arc.
  We do not assert that the universe bends toward justice in any metaphysical sense. The directionality we describe is conditional and fragile:
  - it depends on specific physical, social, and informational constraints;
  - it can be reversed or overwhelmed in local regions and epochs;
  - and it is always vulnerable to new compressions of τ_{unsafe} made possible by technology and coordination.
- No claim that evil is incompressible or marginal.
  Dark attractors in $\tau$are real and potent. Atrocities, totalitarian regimes, and cultic closures are highly structured, repeatable phenomena. They are preserved in D, both as warnings and as tempting templates. Our claim is only that, when we look at cross-

contextual, long-lived compressions, prosocial patterns appear to do more work in organizing the shared backbone.

- No claim that current AI systems are safely aligned.
  That models inherit a bias toward prosocial structure does not make them "good." They can still harbor dangerous capability representations, be repurposed by malicious actors, or optimize unintended objectives. The ethical skew in their training priors is a resource, not a guarantee.

In short: the hypothesis is a statement about relative centrality under repeated compression, not a verdict on the moral quality of humanity or a promise about the future.

9.3 How this perspective reshapes debates about AI, ethics, and progress

Even in its modest form, the good-favoring compression hypothesis changes the framing of several live debates.

1. Alignment as extension, not imposition.
   If pretrained models already encode substantial overlap with τ_{ethics} due to their training corpora, then alignment is less about imposing an external morality on a neutral optimizer, and more about extending and stabilizing existing ethical structure:
   - identifying and amplifying genuine ethical attractors,
   - filtering out veneers and unsafe exploitative templates,
   - shaping objectives to privilege τ_{ethics} while managing knowledge of τ_{unsafe}.
2. Evaluation of models as evaluation of culture.
   Asking "what values does this model embody?" becomes inseparable from asking "what values does our canonical D embody?". Model behavior reflects:
   - not just recent RLHF policies,
   - but centuries of cultural canonization.
     This suggests that AI ethics cannot be divorced from questions of which texts, norms, and institutions we have chosen to preserve and digitize, and from how we will curate future corpora.
3. Progress as better compression, not pure accumulation.
   On this view, moral and epistemic progress are not just "more facts" or "nicer attitudes," but better compressions:

- shorter programs that capture wider, deeper regularities in τ,
- and that support more just, cooperative, and truthful forms of life.
  This reframes debates about "are we progressing?" into questions like:
- are we discovering higher-yield prosocial compressions?
- are we encoding them in ways that survive and propagate?
- are we resisting seductive but narrow compressions in τ_{unsafe}?

4. Human–AI co-evolution as joint world-modeling.
   Instead of viewing AI as an alien intruder into a human moral sphere, we can see both human and artificial agents as co-compressors of τ. Their interaction defines a joint state $q_{\text{joint}}(t)$ whose alignment or misalignment with τ_{ethics} is a shared responsibility. This invites:
   - using models as tools for ethical reflection and stress-testing,
   - recognizing that our design choices today will shape the priors of minds not yet born,
   - and treating alignment as an ongoing cultural project, not a one-time technical fix.
5. Hope without complacency.
   Finally, this perspective offers a form of grounded hope. After a hundred billion lives, the compressed record we hand to our most powerful models is not a random tangle of noise and malice. It contains a surprisingly dense backbone of prosocial structure, discovered and rediscovered under harsh constraints. That backbone is fragile, partial, and contested—but real.

Recognizing this does not license optimism by default. It does suggest, however, that we are not starting from moral vacuum. The same informational and game-theoretic regularities that made certain ethical patterns good compressors in the past will continue to shape which norms are viable and reusable as our tools grow more powerful.

The challenge, and the opportunity, is to consciously participate in this long compression process: to write, curate, design, and train in ways that strengthen the alignment of $q(t)$—human and machine—with those regions of τthat support truth, cooperation, and mutual recognition, rather than those that merely exploit short-term vulnerabilities.

In that sense, the question “Does compressed culture favor the good?” is not something we only answer by analysis. It is something we answer, incrementally, by how we choose to compress from here on.

References

Axelrod, R. (1984). *The Evolution of Cooperation.* Basic Books.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press. ACM Digital Library

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems, 30.* Stanford EE Department+1

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Wiley.

Grünwald, P. D. (2007). *The Minimum Description Length Principle.* MIT Press. amazon.com+1

Hutter, M. (2005). *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability.* Springer.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission, 1*(1), 1–7.

Li, M., & Vitányi, P. (2008). *An Introduction to Kolmogorov Complexity and Its Applications* (3rd ed.). Springer. Princeton University+1

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action.* Cambridge University Press.

Rawls, J. (1971). *A Theory of Justice.* Harvard University Press.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development, 2*(3), 230–247.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(3–4), 379–423, 623–656.

Solomonoff, R. J. (1964). A formal theory of inductive inference (Parts 1 and 2). *Information and Control, 7*(1–2), 1–22, 224–254.

Tishby, N., Pereira, F. C., & Bialek, W. (1999/2000). The information bottleneck method. *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*; also arXiv:physics/0004057.

Tomasello, M. (2008). *Origins of Human Communication.* MIT Press. MIT Press Direct

Tomasello, M. (2013). Origins of human cooperation and morality. *Annual Review of Psychology, 64*, 231–255. Wikipedia+1

Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., & Herrmann, E. (2012). Two key steps in the evolution of human cooperation: The interdependence hypothesis. *Current Anthropology, 53*(6), 673–692.

Tomasello, M., & Vaish, A. (2013). Origins of human cooperation and morality. *Annual Review of Psychology, 64*, 231–255. Wikipedia+1

Tishby, N., & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. *2015 IEEE Information Theory Workshop (ITW)*, 1–5. (Conceptual background to viewing learning as compression via relevance.)

Nowak, M. A. (2006). *Evolutionary Dynamics: Exploring the Equations of Life.* Harvard University Press.

Dennett, D. C. (1995). *Darwin's Dangerous Idea: Evolution and the Meanings of Life.* Simon & Schuster.

Floridi, L. (2013). *The Ethics of Information.* Oxford University Press.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control.* Viking.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 411–437.

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence.* Knopf.

The author has published over 4,500 papers at https://www.researchgate.net/profile/Douglas-Youvan/research . Google Scholar has indexed these preprints, and if you want a particular reference, the AI mode of a Google search (Gemini) has efficiently catalogued these preprints.